Ishan Madhuranga

+353 830 242 827 | contact@excellenttorch.com | LinkedIn | Github | Portfolio
Dublin, Ireland

Professional Profile

AI/Machine Learning Engineer with 4+ years of work experience in delivering end-to-end AI solutions. Skilled in Python, C++, and AWS/Azure Cloud Deployment, with a track record of developing optimized AI Models.

Education

MSc (Hons) Artificial Intelligence — **Jan. 2023 – Jan. 2025**
Dublin Business School — *Dublin, Ireland*

BSc (Hons) Cyber Security Management — **April. 2021 – April. 2022**
Solent University — *Southampton, UK*

Experience

Lead AI Gameplay Engineer — **Sept. 2024 – Jun. 2025**
Sleepless Inc. — *Dublin, Ireland*

- Implemented an AI Gameplay System along with NPC AI behaviours through C++ using Unreal Engine 5.
- Improved game efficiency by profiling performance bottlenecks, optimizing AI pathfinding, and refining blueprint execution with C++, resulting in a 60% reduction in runtime.
- Helped team members to refactor code to maintain clean code.
- Collaborated in Agile sprints and followed SDLC practices to iteratively develop through GIT/Diversion.
- Staying up to date with the latest Unreal Engine advancements and best practices.

Artificial Intelligence Engineer — **Nov. 2021 – Apr. 2023**
British College of Applied Studies — *Colombo, Sri Lanka*

- Implemented production ready projects involving ML and Deep/Reinforcement Learning Models through automation and improving model accuracy and evaluation of trained models using analytics like Tableau, RapidMiner.
- Built and deployed RESTful APIs using FastAPI and Flask, reducing system runtime.
- Deployed models on AWS and Azure, ensuring scalable experimentation and delivery with proper authorization frameworks like OAuth.
- Took responsibilities for networking tasks to avoid security risks.
- Documented all development processes, including model training, evaluation, and deployment strategies, in line with academic and industry standards.

Freelance AI/ML Developer — **Aug. 2019 – Feb. 2025**
Fiverr, Upwork & Direct Contracts — *(Remote) Sri Lanka*

- Delivered AI and ML solutions to clients via Fiverr, Upwork & Direct Contracts
- Collaborated with clients to understand business requirements.
- Built and deployed models for deep learning, computer vision, and data analytics tasks.
- Conducted model optimization and integration for real time applications and IOT/Client based devices
- Provided Data Pre-processing and Model Deployment using Amazon ECS/ Azure Kubernetes.

Projects

AI Research Assistant with RAG & LLM Integration | Python, React, TypeScript, LangChain

- Implemented a LLM based Research Assistant to help students and researchers quickly extract insights and solutions from academic papers, reducing the need to read lengthy documents.
- Built an end-to-end AI solution with a RAG pipeline (LangChain + FAISS) and integrated LLMs for multi-turn reasoning, summarization, and document analysis.
- Optimized prompts for better accuracy and to avoid model hallucinations.
- Developed and deployed a FastAPI backend with a React + TypeScript dashboard.

Fast Neural Style Transfer In a 3D Environment in Real Time | Python, PyTorch, ONNX

- Developed and trained a custom Transformer model for Fast Neural Style Transfer for 3D environments in unreal engine 5.
- Implemented intelligent NPC behaviour using ConvAI for conversations with a custom fine tuned BART model with a RAG pipeline running on the backend to give unique personalities.
- Achieved low latency rendering while maintaining visual quality with Custom Depth Filters in Unreal Engine.
- Quantized the ONNX Models for better performance.

Optimized Inference Engine for ARM/Mobile Devices | C++, ONNX, TensorRT

- Optimized deep learning models for real time performance on ARM/Mobile devices by converting PyTorch models to ONNX and accelerating inference with TensorRT.
- Applied quantization and pruning techniques to shrink model size and cut latency, achieving 20%-30% increase in speed of execution in NVIDIA Jetson Nano hardware.
- Packaged the solution in Docker by ensuring reliability and scalability for production like deployments.

Real Time One Step Latent Diffusion Model for Style Transfer | Python, PyTorch

- Adapted SDXS-512-DreamShaper, a real time one step latent diffusion model, for neural style transfer, achieving low latency high resolution image generation from text prompts.
- Implemented fine tuning and evaluation pipeline in Python/PyTorch, optimizing inference efficiency while balancing fidelity and style consistency for real time applications using Web Cam.

Sentiment Analysis of Tweets: RNNs vs Transformers | Python, PyTorch, Hugging Face, scikit-learn

- Conducted a comparative study of RNN (LSTM) and Transformer models for tweet sentiment classification.
- Implemented models from scratch, pretraining + fine-tuning, and benchmarked against Hugging Face default pipeline.
- Achieved 83% accuracy with LSTM vs 82% with Transformer, while Hugging Face default model underperformed by 60%.

Movie Recommendation System | Python, Streamlit, Scikit-learn, Pandas, NumPy

- Designed and implemented a movie recommender system using the MovieLens dataset (45k movies, 26M ratings reduced version).
- Developed four recommendation methods. Non-Personalized & Personalized Content-Based, SVD Latent Factor model, and Neighborhood + User Based Collaborative Filtering.
- Evaluated model performances using metrics such as RMSE, Confusion Matrix, precision, and recall.
- Built an interactive Streamlit web app to test and compare recommendation algorithms.

Autonomous Self-Driving Car (Prototype) | Python, NVIDIA Jetson Nano

- Implemented an autonomous self-driving car using a custom CNN model trained from scratch deployed on a single board computer.
- Real time inference from a single RGB camera for dynamic navigation decisions using white lanes.
- Implemented Real time Motor RPM Adjustments based on Inference Results.

Skills

- **Languages**: Python, C++, C#, HTML/CSS, JavaScript/TypeScript, SQL, MATLAB
- **TechStack**: Git, HuggingFace, Tableau, Spark, Amazon SageMaker, Azure DevOps, MLDocker, Kubernetes, MongoDB, PostgreSQL, Amazon RDS, Jupyter Notebook, Neo4j, CUDA, cuDNN
- **Libs/Frameworks**: PyTorch, Tensorflow, Scikit-learn, Keras, NumPy, Pandas, OnnxRuntime, FastAPI, Flask, React, FastAPI, LangChain, LlamaIndex, Matplotlib
- **AI/ML**: Deep Learning, NLP, Computer Vision, Reinforcement Learning, Transformer, Generative AI, RNN, LSTM, MLOps, Agentic AI, Meta Learning
- **Methodologies**: Agile, Scrum, SDLC, AI Ethics & Bias Mitigation